

**The Knowledge Bank at The Ohio State University**

**Article Title:** Definition of Textological Data for Coptic Texts

**Article Author:** Orlandi, Tito

**Issue Date:** December 1987

**Publisher:** William R. Veder, Slavisch Seminarium, Universiteit van Amsterdam,  
Postbus 19188, 1000 GD Amsterdam (Holland)

**Citation:** *Polata Knigopisnaia: an Information Bulletin Devoted to the Study of Early Slavic Books, Texts and Literatures* 17-18 (December 1987): 96-105.

**Appears in:**

**Community:** [Hilandar Research Library](#)

**Sub-Community:** [Polata Knigopisnaia](#)

**Collection:** [Polata Knigopisnaia: Volume 17-18 \(December 1987\)](#)

## DEFINITION OF TEXTOLOGICAL DATA FOR COPTIC TEXTS

TITO ORLANDI  
Corpus dei manoscritti copti letterari  
P.le Aldo Moro, 5  
00185 ROMA

When I received a kind invitation to this congress, for a while I was in doubt as to where I should insert my paper. In fact, the subjects of all the sections are more or less interrelated with one another, and the scholars interested in mediaeval manuscripts and in their automatic treatment work more or less in all of them. At last I decided for this section, because it seems to me that the "definition of textological data" corresponds to what the research group in Rome with whom I collaborate, namely "Informatica e discipline umanistiche", calls "codifica" (encoding), and we believe that it represents the root of the relation between our disciplines and Information Technology ("informatica"). On this it seems to depend all possibility to obtain many different results in the research on the manuscripts and the texts which they contain.

On the other hand, the process of encoding whichever material, be it textological, codicological, or group of data, is the most simple to define and organize, from the technical point of view. It consist exclusively in the accurate application of the well known principle of the CORRISPONDENZA BIUNIVOCA, viz.: that each phenomenon in the set of phenomena subject to the encoding (in other words, the part of "world" which the encoder takes into consideration) must have one and only one symbol to express it, and viceversa, that that symbol must have no other meaning.

It is true that more often than not the scholars in human sciences tend to forget that principle, or not to apply it consistently, and the philologists have a good record for that (before the spread of computers and after), possibly due to the fact that language and writing are two relevant examples of very imperfect coding systems. It is also true that some minor problems would remain, eg. the use of a symbol to modify the meaning of other symbols (which, in my opinion, should be avoided wherever possible). But, on the whole, provided the principle is recognized and good will is devoted to apply it correctly, no important technical problems remain to be discussed.

Somebody might draw the attention, in this regard, to the relation and interference between different keyboards, different video displays, different printers; and to the

III phase: semantic analysis (concordance, translation, etc.).

In the first phase the manuscript is encoded in the most faithful way. The first phase is the fundamental one and the editor must limit himself in the operation of encoding without any intervention of interpretative editorial or explaining kind. This of course for what is possible because every such operation involves in part a subjective intervention. The results should ideally substitute the manuscript and make useless the recourse to the manuscript except for verifying errors of reading. In this phase the eventual kind of printing or communication or analysis should remain out of the scope. The editor has only to choose what is to be encoded and the way to encode it. The faithfulness to the manuscript is only partially related to the future reproduction because in this phase we do not take yet into consideration the problems of printing. We have done our proposals for what regards the elements to encode. We have acted according to the practical possibilities given by the keyboards as they are in the market in order to obtain an easy operation of input and have as a result a standard ASCII file, excluding the use of the function codes and also the double codes, and we rely on the standard USASCII keyboard. The number of elements that we can encode is therefore 95 including the space. With this number it is possible to obtain a good encoding of a normal manuscript. The elements that we have enucleated are as follows:

The letters of the Coptic alphabet, each with eventual superlinear stroke ( $33 + 33 = 66$  signs).

The letters "iota" and "ypsilon" with the diaeresis.

5 punctuation signs.

Change of page, line, column.

Capital letter in the margin.

5 signs for special superlinear strokes.

Illegible letter; letter in a physical lacune.

Remarks by the editor.

Original page numeration.

PROBLEMS: The punctuation cannot be encoded according to the physical appearance of the signs, because the scribes tend to be inconsistent. The editor should declare what seems to be the system of the manuscript, and then interpret the signs according to the intention of the scribe.

We have eliminated the category of the letters "not quite readable, but presumable", because such interventions by the editor are better placed in the "second phase" (cp. below). Also the editor should refrain from filling any lacuna, in

the field of Coptic manuscripts and texts, beginning with some historical information outside the work in our Corpus dei Manoscritti Copti Letterari. We mention first the enterprise of the Nag Hammadi project, done at the Institute for Antiquity and Christianity of the Claremont (California) Graduate School with the well known Ibycus system. The Ibycus was first conceived by David Packard for Latin and Greek texts. Its encoding system permits some degrees of textual analysis, and above all the exit on photocomposer, which produces a very good printed text. Its main drawback is that it is too much "printing-oriented", thus e.g. providing a code to put missing or uncertain letters within brackets, rather than signalling what really IS in the manuscript. Furthermore, it provides a code to signal whether a letter has some superlinear stroke or other such peculiarities, instead of a code individuating the "letter with superlinear stroke" (if this is considered an individual phenomenon), or else a code to indicate the superlinear stroke and its position. It also tends not to distinguish textual from codicological phenomena (cp. above). What the Claremont enterprise has produced up to now are beautiful printed editions (some Nag Hammadi Texts and also a dictionary), but not other results. On the same line we may put the Princeton enterprise for Old Testament, which uses the same Ibycus system, but has not yet yielded practical results.

On a different level we should mention the tools which simply provide for the possibility to print Coptic fonts. Some of them are meant to accompany a word-processor installed on a Personal Computer machine, instructing the screen and the printer (Academic Font; Toolbox for Languages; Lettrix; etc.); others provide the fonts for photo-typesetting, and are professional oriented for the typography. All this is out of our scope, here.

For our system in the Corpus dei Manoscritti Copti Letterari we have worked on the basis of the following principles:

To encourage the collaboration of the scholars;

to have files fully portable on all kind of machines, great and small, supporting all kind of printers;

- to use a delimited range of program types, which may be individually different: editors, word-processors, text-formatters, data-bases, concordance programs.

The manuscripts are encoded in one file, which contains all the indications which may be useful in future, but are selected at different phases in order to be submitted to different procedures.

I phase: diplomatic reproduction.

II phase: edition of the text, normalized in orthography, and divided into paragraphs according to the modern taste.

also in the operation of transcoding the written text, the elements of subjective interpretation by the scholar are present, and he (like any text editor) must assume his responsibility in the choices, and of course declare them as clearly as possible.

Much more subjective is the operation of encoding the "graphic organization" of the manuscript. As we have said, the primary interest, also here, is represented by the text; therefore what the scholar should do first, is to discern the relation between the graphic organization and the meaning of (parts of) the text, and choose those phenomena to be encoded, which are "meaningful" in very broad sense. One way to do it is, eg., to extrapolate a "regular" graphic organization in that particular manuscript (columns, lines, margins, etc.), in which are obviously collocated the "normal" parts of the text. (For "normal" parts we simply assume those letters which are kept within the boundaries of the "regular" graphic organization). The encoder will signal with appropriate codes wherever some group of letters is out of the "regular" place, or some part of the "regular" place is not filled with letters.

As the graphic organization which one assumes at the beginning is somewhat imaginary, that is, it is an imaginary regularization of the actual graphic organization of each page of the manuscript, the alterations which it suffers by the actual position of the scripture in the manuscript are infinite. For instance, it is often impossible to establish whether a group of letter are in "interlineo" from its physical position, because a scribe may write some letters a bit up, only by chance. This is why the operation of encoding presupposes the subjective interpretation of the editor, who will act on the basis of his apprehension of the relation between the graphic organization and the meaning of the text (including the groups of letters and marks accompanying the text proper).

Finally we would add some remarks about the correctness of the encoding, that is, how we can judge whether a manuscript is encoded correctly or not. From this point of view one should consider that the aim of encoding a manuscript is not simply that of later obtaining a faithful reproduction on different support (e.g. a screen) or by means of electronic processing (e.g. the print from a computer printer). The aim is also that of producing various kinds of textual or codicological analysis. Therefore the correctness of the encoding (that is, of the choice of the phenomena to be encoded) depends on the final product which is to be obtained, where this product should not be seen only as a book or a traditional edition of the text. Thus we may judge the correctness from the possibility that the encoding gives to reach the aims which we want, or others may possibly want in future. Nothing else is required, because the choice of the sign within the code is in itself irrelevant, on account of the possibility to reshape it automatically.

We come now to the practical application of this theory in

between these single elements which will permit the use of a logical information retrieval language, whichever it is. (In this sense, also a concordance program or a lexical analyzer.)

The pitfall in this operation (as I happened to notice in many cases) is that scholars tend to confuse between encoding and transcoding. I call an encoding just the operation alluded to above; a transcoding is more simply the encoding done on already encoded material. In this case, we have simply to substitute each sign of an alphabet (in broad sense) with that of another one, employed because it may be "written" on a different support (eg. the Morse alphabet). If we have this in mind, it is easy to understand that the "text" is simply to be transcribed, because it is already encoded in the written alphabet or scripture (though here also problem arise, due to the imperfection of the alphabets as such). On the contrary the visual organization of the text and the material organization of the codex are the object of a true, "primitive", encoding process. Therefore, the problems pertaining to the two operations should be accurately kept distinct, even if the result should be unitarian, viz. the production of ONE file of encoded information, because in this paper it is assumed that the interest of the scholars is ultimately centered on the text.

We shall consider first the transcoding of the text. The problems here arise from the fact that we find the text already encoded in the manuscript, but by means of a peculiar alphabet, in the sense that its signs (the "letters" and other marks) are to be recognized only in part from their material form or substance, and in part from the relation between that form and (a) the general meaning of the text; (b) the position in the page (cp. the page numbers; the titles; the glosses; etc.).

It is well known to every palaeographer that the single letter, with its different forms, due to the skill of the scribe, but also the various conditions in which he works, is recognized only in part because of its form, and in part from the fact that, given the context, that letter "must" be that letter. (Attention! I do not allude to the possibility of confusion between two letters; that is a different problem, which cannot be solved by the context. I allude to the often peculiar forms of one single letter).

There is another problem. Each letter has different meanings, that is, it refers to more than one "single phenomenon", and those meanings depend sometimes on its form (capitals etc.), sometimes on its position (numeration etc.). The scholar should decide whether: (a) to propose a true and simple transcodification, by which the new signs acquire the same duplicity of the old ones; (b) to propose a kind of ameliorated transcodification, in which the same sign is transcribed in different ways according to its different meanings, thereby correcting (for what is possible) the incorrect encoding of the ancient, traditional, scribes. In any case we want to stress that

various systems incessantly proposed to obtain comfortable ways to input texts and to read them. We all know well such problems, and that their solution is probably left to the technological progress and the skill of some engineers, not to particularly brilliant ideas by the scholars in the humanities. Therefore I would not take them too seriously, though I realize that we must always try and improve the machines with which we work.

So we are left totally on the other side of the question. The scholar has to identify very carefully the phenomena to encode (what is properly called in the "Call for papers": discretization of continua) within the material which forms the subject of his study. A first consequence is that all discussion on the standard for the different languages and purposes should center on this matter, not on the choice of the symbols (which, in any case, may be easily translated by means of elementary programs of translation from one code to another).

More important is that scholars do not seem to realize how difficult it is to carry on properly the task of identifying the phenomena to encode. In a sense, to this task were devoted all scholars since the beginning of their sciences; but this is even more deceiving, because the stringent consistency of the machine has shown how different is the treatment of data to be communicated by means of natural language in a monograph, from the one devised in order that information (in broad sense) may be retrieved by means of a computer.

But there is more, especially if we turn to the particular branch of codicology and textology. The first idea to be firmly kept in mind is that the manuscript, from one side, and the text, from the other, are two entirely different things, having a sort of dialectical relationship between themselves. On this relationship we shall deal later; now we start from the manuscript, considering it as a material artifact, and noting that infinite are the phenomena in it, which might interest scholars, either in the time when they are working, or later. Every particularity in its construction, every sign or mark put in by the scribes (and correctors), and their relative position, may eventually prove important.

From this point of view, the only satisfactory way to store all this in a magnetic memory is the analogical, not the digital reproduction, that is, a continuous and not a discrete one (photo, videotape, videodisk, etc.). Or, to be more precise, we do not see today a way to store such information so that its discrete elements (because after all discrete they are, even in a photo or in a video-disk) may be treated as logical elements for information retrieval.

The "real digital" storage is obtained, just as we said, through a first step done by the scholar, never by the machine, consisting in the individuation of the elements to encode. It is the objective logical and factual relations

this phase.

In principle, the editor should not even encode spaces between words, unless they are in the manuscript. But it is not harmful to add such spaces, and it is advisable to do so, in order to spare time in the "second phase", when the division between words (or rather grammatical entities) must be done.

II PHASE. In this phase the editor leaves aside the point of view dedicated to the manuscript, and assumes that of the text itself as an entity. Therefore another file is formed, derived from the fundamental one, where the codes for the elements proper to the manuscript (line division, interspersed blanks, column and page division, abbreviations, etc.) are eliminated, normally by automatic processing.

We are also in favour of the elimination of the code for the superlinear stroke (except for special texts), because its use had to do with the ancient reading habits, rather than the meaning of the text, and it can be substituted by the separation of words.

The editor will now fill (wherever possible) the lacunae; will change or insert punctuation, in order to normalize the text according to logical paragraphs and sentences; and also will normalize orthography, though this is a point still very debatable.

From this file the editor will obtain, through automatic processing, a formatted edition and the concordance.

III PHASE. This phase cannot be defined precisely as the previous ones. It is implemented when there are more manuscripts for the text, each of which has been previously treated as stated above. The aim is to produce the critical edition, through the comparison of the different readings of the manuscripts, and make the lexical and semantic analysis for which appropriate programs may be prepared. Also some kind of automatic or semi-automatic translation might be envisaged.



ΞΕΝΘΑΔΕ ΝΤΕ  
 ΠΕΝΔΕΡΙΤ ΝΙ  
 ΩΤ ΔΠΑ ΠΑΥΛΕ  
 ΕΤΒΕ ΤΑΔΑΚΡΗ  
 ΟΙΟ· ΞΕΝ ΟΥΕΙΡΗ  
 ΝΗ· ΞΞΔΗΝ:-  
 ΞΑΘΗ ΜΕΝ Π  
 ΞΩΒ ΝΙΩ ΔΡΙ  
 ΞΟΤΕ· ΝΞΗΤΗ  
 ΜΠΝΟΥΤΕ· ΠΓ  
 ΞΑΡΕΞ ΕΝΕΧΕΗ  
 ΤΟΛΗ:-  
 ΝΙΕΝΤΟΛΗ ΒΕ  
 ΜΠΝΟΥΤΕ ΝΕ  
 ΝΔΙ· ΠΞΩΗΡΕ  
 ΕΚΝΔΞΑΡΕΞ ΕΡΟΟΥ  
 ΜΙΟΩΕ ΞΕΝ ΟΥ  
 ΜΠΤΖΗΚΕ· ΜΠ  
 ΟΥΒΙΟΟ· ΕΒΑΔΩΟΥ  
 ΔΥΩ ΟΥΜΠΤΞΤ  
 ΡΟΟΥΨ· ΜΠ ΟΥ  
 ΒΡΩΞ· ΤΞΡΕΚ  
 ΒΩ ΕΥΜΟΤΕΝ:  
 ΕΠΕΙΔΗ ΓΑΡ·  
 ΠΑΔΟΝ ΤΜΠΤ  
 ΞΗΚΕ ΜΠ ΠΕΒ  
 ΡΩΞ· ΠΕΤ\_Ω  
 ΞΠ ΕΒΟΛ ΞΜ Π  
 ΠΑΘΟΟ:-  
 ΔΥΩ ΤΜΠΤΞΤ

4Ε  
 ΡΟΟΥΨ ΠΠΞΠ· ΠΕΤ  
 ΤΟΥΞΟ· ΜΠΡΩΜΕ  
 ΜΠΕΡΨΗΝΕ· ΝΟΟ  
 ΞΑΙ· ΝΒΙΝΕΡΩΒ  
 ΤΑΡΕΚΕΨΕ ΞΕ Π  
 ΚΗ ΞΡΞΙ ΝΔΨ Π  
 ΞΕ:-  
 ΔΥΩ ΜΠΕΡΔΩΟΪ  
 ΞΠ ΝΕΚΜΕΥΕ  
 ΤΑΡΕΚΒΩ ΕΚΟΒ  
 ΡΑΞΠ· ΔΥΩ ΝΓ  
 ΤΕΞΤΩΞ ΔΝ:-  
 ΤΕΝΟΥ ΒΕ ΠΔΨ===  
 ΒΩΨΤ ΕΝΕΤ==  
 ΔΔΒ· ΤΗΡΟΥ· ==  
 ΝΔΥ· ΞΕ ΝΤΔ\_  
 Π ΟΥΔ· ΝΟΥΩΤ  
 ΜΠ ΠΝΟΥΤΕ· ΝΔΨ  
 ΝΞΕ:-  
 ΝΤΔΥΜΟΟΩΕ ΓΑΡ  
 ΜΠ ΟΥΜΠΤΩΜ  
 ΜΟ· ΜΠ ΟΥΜΠΤ  
 ΞΗΚΕ· ΜΠ ΟΥΒ  
 ΡΩΞ· ΜΠ ΟΥΤΕ (sic)  
 ΕΒΟΛ· ΝΔΞΡΕΝ ΟΥ  
 ΟΝ ΝΙΩ· ΞΔΝ  
 ΤΟΥΞΡΟ· ΕΠΞ.ΝΔΙ  
 ΚΙΜΕΝΟΟ:-  
 ΕΥΟΡΜ· ΞΝ ΝΕΞΔΙΕ  
 ΜΠ ΝΕΤΟΥ ΜΠ  
 ΝΕ\_· ΜΠ ΝΕΨ  
 ΚΟΛ ΜΠΚΑΞ  
 ΕΥΡ ΒΡΩΞ· ΕΙΘΑΙΒΕ

1. ΖΑΘΗ ΜΕΝ ΝΖΩΒ ΝΙΜ ΔΡΙ ΖΟΤΕ ΝΖΗΤΥ ΜΠΝΟΥΤΕ 46  
ΝΓΖΔΡΕΖ ΕΝΕΨΕΝΤΟΛΗ:

2. ΝΙΕΝΤΟΛΗ ΒΕ ΜΠΝΟΥΤΕ ΝΕ ΝΔΙ ΠΔΨΗΡΕ. ΕΚΝΑΖΔΡ-  
ΕΖ ΕΡΟΟΥ. ΜΟΟΨΕ ΖΕΝ ΟΥΜΝΤΖΗΚΕ ΜΝ ΟΥΒΙΟC ΕΨΑΖΔCΩΟΥ  
ΔΥΩ ΟΥΜΝΤΑΤΡΟΟΥΨ ΜΝ ΟΥΒΡΨΖ ΤΑΡΕΚΒΩ ΕΚΜΟΤΕΝ:

3. ΕΠΕΙΔΗ ΓΔΡ ΠΔCΟΝ ΤΜΝΤΖΗΚΕ ΜΝ ΠΕΒΡΨΖ  
ΠΕΤ(Β)ΩΔΠ ΕΒΟΛ ΖΜ ΠΠΔΘΟC ΔΥΩ ΤΜΝΤΑΤ#ΡΟΟΥΨ ΝΝΔΙ 2  
ΠΕΤΤΟΥΔΟ ΜΠΡΨΜΕ:

4. ΜΠΕΡΨΙΝΕ ΝCΔ ΖΔΖ ΝΒΙΝΕCΠΖΩΒ ΤΑΡΕΚΕΙΜΕ ΔΕ  
ΚΚΗ ΖΡΔΙ ΝΔΨ ΝΖΕ ΔΥΩ ΜΠΕΡΔΨ(Δ)Ι ΖΝ ΝΕΚΜΕ(Ε)ΥΕ ΤΔ-  
ΡΕΚΒΩ ΕΚCΒΡΑΖΤ ΔΥΩ ΝΓΤΕΖΤΨΖ ΔΝ:

5. ΤΕΝΟΥ ΒΕ ΠΔΨ(ΗΡΕ) ΕΨΨΤ ΕΝΕΤ(ΟΥ)ΔΔΒ ΤΗΡΟΥ  
(ΝΓ)ΝΔΥ ΔΕ ΝΤΑ(Υ)Ρ ΟΥΔ ΝΟΥΨΤ ΜΝ ΠΝΟΥΤΕ ΝΔΨ ΝΖΕ:

6. ΝΤΔΥΜΟΟΨΕ ΓΔΡ ΜΝ ΟΥΜΝΤΨΜΜΟ ΜΝ ΟΥΜΝΤΖΗΚΕ ΜΝ  
ΟΥΒΡΨΖ ΜΝ ΟΥCΤ(Ο) ΕΒΟΛ ΝΔΖΡΕΝ ΟΥΟΝ ΝΙΜ ΨΔΝΤΟΥΔΡΟ  
ΕΠΑΝ(Γ)ΙΚ(Ε)ΙΜΕΝΟC:

7. ΕΥCΟΡΜ ΖΝ ΝΕΔΔΙΕ ΜΝ ΝΕΤΟΟΥ ΜΝ ΝΕ(ΙΑ) ΜΝ  
ΠΕΨΚΟΔ ΜΠΚΔΖ ΕΥΡ ΒΡΨΖ ΕΥΘΛΙΒΕ # ΕΥΜΟΚΖ ΝΔΙ ΕΤΠΠΕ-ΟC- 46  
CΨC ΜΠΨΔ ΔΝ ΜΜΟΟΥ:

8. ΔΡΙ ΠΜΕ(Ε)ΥΕ ΟΝ ΜΠΕΝCΨΤΗΡ ΔΕ ΝΓΔΗΕΙ ΖΝ  
ΟΥΜΝΤΖΗΚΕ ΔΨΚΩ ΝΔΝ ΖΡΔΙ ΝΝΕΨ(Ζ)ΙΟΟΥΕ ΔΕΚΔC (ΕΝ)ΝΔ-

```

}MONB.GU
}TITO ORLANDI   Revisione 25 febbraio 1987
}Microfilm dall'archivio CMCL
}Particolarit  grafiche: la I } scritta sempre col punto
}sopra; nella trascrizione } notata solo la dieresi.
}I paragrafi sono accompagnati da coronis in colonna
}sinistra, da obelos in colonna destra. La sopralinea
}> puntiforme, e varia di collocazione.
}=====
?WE?
}>titolo con ornamenti
<EN'+A+E nTE !PEN'MERIT N'I!VT APA PAULE !ETBE T'DIAKRH!SIS.
<EN OU'EIRH!NH. \AMHN:4 I'
}>fine titolo   Il titolo appartiene probabilmente al testo precedente
}=====
}>De paupertate
}-----
<<AJH MEN n'!\VB NIM ARI !OTE$ n'HTW !mP'NOUTE. nG'!\ARE<
ENEW'EN!TOLH:4444 !\NI'ENTOLH 'E 'mP'NOUTE NE 'NAI. !PA'+HRE !
EKNA'\ARE< EROOU !\MOO+E <EN OU'!MNT<HKE. !MN !OU'BIOS. EB'ASVOU '
AUV OU'MNTAT!ROOU+. !MN OU'!RV'. !TAREK'!V EK'MOTEN: !EPEIDH
GAR$ !PA'SON T'MNT!<HKE MN PE'!RV<$ PET'-V!*P EBOL <M P'!PAJOS:444 I'
(AUV T'MNTAT#ROOU+ n'NAI. PET'!TOU*O. mp'RVME !\MPER'+INE. nSA !
<A'$ n'INER:VB !TAREK'EIME *E K'!KH <RAI N'A+ n'!<E:444 !!!!!
(AUV MPER'A+O+ ! n NEN'!SEUE !\TAREK'!V EK'S'!RA<t. AUV NG'!TE<TV.
AN:44 !<TENOU 'E PA'+=== !'V+T ENET'==!AAB$ THROU. ==!NAU. *E
NTA='!r OUA$ N'OUVT !MN P NOUTE$ N'A+ !N'<E:444 !!!!!
<NTAU'MOO+E GAR !MN OU !MNT+M'MO. MN OU'MNT!<HKE. !MN OU'!RV<. I
MN OU'STEB 'EBOL. NA BEN OU'ON NIM. +AN!TOU'*RO$ eP'ANDI!KIMENDS:444 I!
<EU'SORM$ <N NE'+AIE 'MN NE'TGOU MN !NE--. MN NE'+!KOL mP'KA<
EU'R 'RV.. EU'JLIBE #

```